LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO

GESTÃO DE SISTEMAS DE INFORMAÇÃO

TRABALHO FINAL DE MESTRADO

TRABALHO DE PROJETO

ANÁLISE E IMPLEMENTAÇÃO DE MELHORIAS DE QUALIDADE DE DADOS NO PROCESSO DE MIGRAÇÃO DA INFORMAÇÃO DE CLIENTES

ANTÓNIA MELICIA DE SOUSA ORDENÃ

OUTUBRO - 2018

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO EM
GESTÃO DE SISTEMAS DE INFORMAÇÃO

TRABALHO FINAL DE MESTRADO
TRABALHO DE PROJETO

ANÁLISE E IMPLEMENTAÇÃO DE MELHORIAS DE QUALIDADE DE DADOS NO PROCESSO DE MIGRAÇÃO DA INFORMAÇÃO DE CLIENTES

ANTÓNIA MELICIA DE SOUSA ORDENÃ

ORIENTAÇÃO:

ENG. ANA MARIA MARQUES RIBEIRO DOS SANTOS LUCAS

OUTUBRO - 2018

Agradecimentos

Sendo este trabalho o culminar de anos de estudo, começo por redigir os agradecimentos às pessoas que me ajudaram a chegar a esta etapa.

Agradeço à minha mãe, Adelaide, pelos esforços que fez por mim e pelos meus irmãos durante estes 24 anos para que nós tenhamos a possibilidade de alcançar as nossas ambições. Agradeço muito a tua força mãe, muito obrigada.

Agradeço aos meus irmãos pelo companheirismo e carinho.

Ao Afonso pelo apoio incondicional, pela paciência, pela força constante para que eu atinja os meus objetivos e principalmente por me mostrar como o mundo a dois é melhor.

À minha tia Nelinha pelo apoio, disponibilidade e representação de força que ajudaram a moldar a pessoa que sou.

Às minhas “Jadas” e aos “Besties” pela amizade incondicional, pelos momentos divertidos e pelo carinho que disponibilizam.

Agradeço também à minha orientadora Eng. Ana Lucas pela orientação que forneceu e tornou este trabalho possível. Acrescento também que a paixão que transparece por esta área de estudos permite aos seus alunos identificar o que realmente lhes interessa.

Por fim agradeço à empresa que trabalho atualmente pela integração neste projeto de qualidade de dados, em especial agradeço ao João pela orientação e apoio que disponibilizou durante e após o desenvolvimento deste projeto.

“It is what we make out of what we have, not what we are given, that separates one person from another.”

Nelson Mandela

Lista de Siglas e Acrónimos

Abreviatura	Descrição
BR	Business Rules
CAE	Classificação de Actividade Económica
DP	Data Profiling
GTQD	Gestão Total de Qualidade de Dados
EEL	Expression Language
LD	Limpeza de Dados
MQD	Melhoria de Qualidade de Dados
NIB	Número de Identificação Bancária
NIF	Número de Identificação Fiscal
PI	Produtos de Informação
PQD	Problemas de Qualidade de Dados
QD	Qualidade de Dados
SI	Sistema de Informação
TDQM	Total Data Quality Management
TQM	Total Quality Management

Resumo

O aumento da quantidade de dados atualmente relevou a importância da qualidade nos dados. Considerando este fator a FinanceQ, empresa do sector financeiro, no âmbito do projeto de migração, reconheceu a importância de melhorar a qualidade dos dados a migrar. Nesse sentido requisitou os serviços da SIGQ, empresa que fornece soluções na área de sistemas de informação, para analisar e implementar medidas de forma a melhorar a respetiva qualidade. O objetivo do projeto centrou-se em três finalidades: analisar a qualidade de dados atual; aplicar medidas de normalização nos dados; e aplicar medidas de enriquecimento nos atributos de morada.

Para cumprir os objetivos definidos a FinanceQ adquiriu o *software SAS Dataflux* e aplicou a metodologia da aplicação composta por três fases: planeamento; ação; e monitorização. Durante o processo de qualidade foram aplicadas técnicas de *data profiling* para analisar os dados e considerou-se a taxonomia de Oliveira et al. (2005) para identificar o tipo de anomalia nos dados. Quanto a melhoria de qualidade de dados seguiu-se a estratégia reativa onde foram aplicadas técnicas de normalização e enriquecimento para solucionar os problemas identificados: valores sem significado; valores a *null*; padrões inadequados para o atributo; erros ortográficos; existência de sinónimos; e valores fora do domínio dos atributos.

Na fase final do projeto foi possível identificar que as técnicas aplicadas permitiram designar corretamente os géneros, reorganizar os números de telefone e validar os padrões de valores, como por exemplo, as datas; as ações de limpeza e correção dos dados permitiram eliminar os valores sem significado e corrigir os erros ortográficos, como por exemplo nos atributos de *email* e *site*; O processo de enriquecimento nas moradas permitiu normalizar os dados e enriquecer os atributos de código postal em 80% dos tuplos.

Na generalidade as técnicas aplicadas tiveram como objetivo melhorar as dimensões de qualidade exatidão, objetividade, completude e consistência.

Palavras-chave: Qualidade de dados, problema de qualidade de dados, melhoria de qualidade de dados, GTQD, *data profiling*, *data enrichment*.

Abstract

The large availability of data that now exists highlights the importance of data quality. Considering this factor FinanceQ, a company that works in the financial sector, acknowledged the importance of improving data quality in their migration project. With this goal in mind, they requested the services of SIGQ, a company that provides information systems solutions, to analyse and implement data quality procedures. The goal of this project centred on three key issues: analysis of the current data quality; normalization of data; and address data enrichment.

To fulfil the defined goals FinanceQ acquired the software SAS Datafl*ux* and applied the SAS Dataflux methodology composed of three steps: planning; action; and monitoring. During the data quality process, data profiling techniques were applied to analyse data and the taxonomy of Oliveira et al. (2005) was considered to identify anomaly types. A data driven strategy was used for quality improvement and the techniques applied were data normalization and data enrichment to solve the identified problems: meaningless values; missing values; inadequate patterns; misspellings; synonymous; and values behind the context.

In the last stage of the project it was possible to verify that the applied techniques allowed for correct designation of the gender fields, reorganization of telephone numbers and identification of measures to validate value patterns; the data cleaning and treatment helped to eliminate meaningless values and correct misspellings, for example in the email and site fields; the data enrichment process of addresses permitted normalisation and enrichment of the postal code fields in 80% of the records.

In general, the goals of the applied techniques were to improve the data quality dimensions accuracy, objectivity, completeness and consistency.

Keywords: Data quality, data quality problem, data quality improvement, TDQM, data profiling, data enrichment.

Índice

Índice de Tabelas

Índice de Figuras

1. Introdução

A era da informação provocou um aumento da quantidade de dados disponíveis, que atualmente podem apresentar diversas formas, são gerados a uma alta velocidade e precisam de ser validados para que sejam usados pelo utilizador. Este paradigma denominado de *Big Data* modificou a concorrência entre as organizações (Baesens et al., 2016; Porter et. al, 2014).

As redes sociais disponibilizadas pelo desenvolvimento da web 2.0 permitiram aos utilizadores a possibilidade de serem criadores de dados (Isaías & Coelho, 2013) . Os objetos inteligentes embebidos de sensores contêm capacidades de monitorização, controlo, otimização e autonomia e produzem dados sobre as suas atividades (Porter et al., 2014). Os dispositivos móveis, através da ligação à internet e das aplicações móveis, geram dados que fornecem informação sobre o comportamento do utilizador (Baesens et al., 2016).

O paradigma *Big Data* está a modificar o processo de tomada de decisão, uma vez que, a análise de dados é cada vez mais utilizada para tomar decisões (Chen, Chiang, & Storey, 2012). No entanto a grande quantidade de dados disponíveis, não significa qualidade. Para que os utilizadores tenham confiança no uso dos dados é importante apurar a respetiva qualidade, para que possam suportar corretamente a tomada de decisões (Davenport, Harris, & Morison, 2010).

A empresa FinanceQ (nome fictício) é uma instituição bancária que vende soluções financeiras e que em 2017 apresentou um volume de vendas de cerca de trinta e cinco milhões de euros. Na sequência da necessidade de inovar a versão da base de dados (AS400) para uma versão mais recente, procederam a implementação de um projeto de migração. No âmbito do respetivo projeto surgiu a necessidade de implementar medidas para melhorar a qualidade dos dados a migrar. O projeto apresentado será focado na fase de intervenção sobre a qualidade de dados.

Tendo como finalidade analisar e implementar medidas de qualidade de dados a FinanceQ requisitou os serviços da SIGQ (nome fictício), empresa onde trabalha a

autora e que fornece soluções no setor dos sistemas de informação, apresentando em 2017 um volume de vendas de cerca de vinte sete milhões de euros.

No âmbito do projeto foram propostos os seguintes objetivos:

- Análise à qualidade dados: análise de anomalias utilizando técnicas de *data profiling*, aplicação de regras de negócio e de validação de tabelas de referência;
- Normalização de dados: limpeza de dados descontextualizados, correção de erros ortográficos e normalização da forma; e
- Enriquecimento de moradas: normalização das moradas e enriquecimento de códigos postais de 7 dígitos utilizando informação pública disponibilizada pelos CTT.

Tendo em consideração o pedido efetuado pela FinanceQ, o objetivo deste projeto centrou-se na análise do estado atual utilizando ferramentas de *data profiling* e aplicações de processos de normalização e enriquecimento nos dados de forma a melhorar a respetiva qualidade.

No âmbito deste trabalho utilizaremos indiscriminadamente as palavras dado e informação, à semelhança de outros autores de qualidade dos dados (Strong, Lee, & Wang, 1997; Wang, 1998).

2. Revisão da Literatura

Uma das áreas onde se notou maior interesse e consciencialização da necessidade de qualidade de dados (QD) foi a área de reporte financeiro, nomeadamente no cálculo dos balanços financeiros (Ballou & Pazer, 1985). Nessa altura as preocupações prendiam-se com a necessidade de haver consistência entre os dados provenientes de diferentes fontes (internas e/ou externas) e na exatidão dos valores (Ballou & Pazer, 1985) .

Com a consciencialização surgiu a necessidade de criar formas de avaliar e melhorar a qualidade de dados (Ballou & Pazer, 1985). Considerando este objetivo tornou-se cada vez mais evidente para a comunidade científica que a forma como o processo de

captação, armazenagem e disponibilização de dados está construído afeta a qualidade de dados, como tal, tornou-se importante considerar o contexto do SI no conceito de QD (Orr, 1998; Strong, Lee, & Wang, 1997; Tayi & Ballou, 1998; Wang & Strong, 1996).

Wang e Strong (1996) identificaram a necessidade de analisar as dimensões de QD para além da reconhecida exatidão. Nesse contexto criaram uma *framework* que identificou as dimensões de QD importantes para que os dados estejam ajustados ao uso do utilizador. O estudo de Wang e Strong (1996) permitiu reconhecer que as dimensões de QD podem ser agrupadas em quatro categorias: intrínseca, contextual, representacional e acessibilidade. A etapa seguinte para a comunidade científica passou pela introdução de modelos que permitissem avaliar o estado atual da QD, considerando as dimensões de QD.

Ballou et al. (1998) construíram um modelo que surgiu da analogia do processo de criação de informação com o processo de fabrico de produtos, que nomearam de modelo de fabrico de informação. O objetivo do modelo apresentado era determinar o tempo, custo, a qualidade e o valor dos produtos de informação de forma a serem valorizados pelos utilizadores finais.

Wang (1998) desenvolveu uma metodologia, designada de gestão total de qualidade de dados (GTQD), que utiliza as dimensões de QD introduzidas por Wang e Strong (1996). O objetivo da metodologia é avaliar o estado atual dos dados e promover processos de melhorias.

Ao longo dos anos foram sendo apresentadas várias definições de QD, apresentando-se algumas na Tabela I.

Tabela I - Definição de Qualidade de Dados

Autor	Definição
(Brodie, 1980, pgs. 246-247)	*"Data quality concerns preserving the meaning of data as percieved by designers and users of a database application. (…)Data quality is a measure of the extent to which a database accurately represents the essential properties of the intended application."*
(Wang, Kon, & Madnick, 1993, pg. 671)	*"(…) we define data quality on this basis. Operationally, we define data quality in terms of data quality parameters and data quality indicators (…).* *(…) A data quality parameter is a qualitative or subjective dimension by which a user evaluates data quality.* *(…) A data quality indicator is a data dimension that provides objective information about the data."*
(Orr, 1998, pg. 67)	*"Data quality is the measure of the agreement between the data views presented by an information system and that same data in the real world."*
(Wand & Wang, 1996, pg. 87)	*"(…) the quality of data depends on the design and production processes involved in generating the data. (…) the notion of data or information quality depends on the actual use of data."*
(Wang & Strong, 1996, pg. 6)	*"(…) we define "data quality" as data that are fit for use by data consumers"*
(Tayi & Ballou, 1998, pg.54)	*"The term "data quality" can best be defined as "fitness for use," which implies the concept of data quality is relative. Thus data with quality considered appropriate for one use may not possess sufficient quality for another use."*

Neste estudo adota-se a definição de Wang e Strong (1996), em que se considera que os dados apresentam qualidade se são adequados à respetiva utilização. Como tal é necessário analisar o tema da qualidade de dados como um conceito multidimensional que depende do contexto em que é utilizado (Tayi & Ballou, 1998).

A título de resumo, conclui-se que o tema de QD esteve desde os anos 80 em constante evolução, em parte devido ao valor que foi dado à qualidade dos produtos com a introdução do modelo TQM (Total Quality Management) (Martinez-Lorente, Dewhurst, & Dale, 1998) , mas principalmente pela entrada da "Era Digital" que trouxe a necessidade de utilização dos dados como forma de diferenciar da concorrência.

2.1. Dimensões de qualidade de dados

A avaliação do nível de qualidade de dados é realizada avaliando as dimensões de QD, que representam características dos dados que são importantes na ótica do utilizador (Wang & Strong, 1996).

A identificação das dimensões de qualidade é realizada tendo em conta três aspetos: as características próprias dos dados; o contexto da respetiva utilização; e a importância dos sistemas de informação (Orr, 1998; Strong et al., 1997; Wand & Wang, 1996; Wang & Strong, 1996).

Considerado o conceito de QD definido por Wang e Strong (1996), as dimensões de QD podem ser agrupadas em quatro categorias, como representado na Tabela II.

Tabela II - Dimensões de Qualidade de Dados

Categoria	Dimensão
Intrínseca	Exatidão, Objetividade, Credibilidade, Reputação
Contextual	Valor acrescentado, Relevância, Oportunidade temporal, Completude, Quantidade apropriada.
Representacional	Interpretabilidade, Compreensibilidade, Representação concisa, Representação consistente.
Acessibilidade	Acessibilidade e Segurança nos acessos.

Fonte: Strong et al. (1997), pg. 104

A categoria intrínseca contém dimensões que analisam se os dados estão corretos e se são objetivos. Numa ótica relacionada com a fonte de dados, esta categoria contém dimensões que permitem avaliar a reputação e a confiança dos dados considerando a respetiva origem.

A categoria contextual contém dimensões relacionadas com o contexto em que os dados são utilizados. A avaliação passa por analisar se os dados estão completos, se permitem auferir valor acrescentado e se são relevantes para o contexto de utilização. O contexto inclui a dimensão que analisa a janela temporal em que os dados são gerados, tal como a dimensão que avalia a quantidade, que deve ser apropriada para que o utilizador tenha informação suficiente e possível de ser analisada.

As dimensões incluídas na categoria representacional têm como objetivo avaliar a interpretabilidade e compreensibilidade dos dados para os utilizadores.

Por último, a categoria acessibilidade contém dimensões que permitem avaliar o acesso aos dados. Neste sentido é necessário que as políticas de acessos não restrinjam o acesso ao utilizador, no entanto, devem ser aplicadas técnicas de segurança para que o utilizador tenha confiança na utilização dos dados.

2.2. Problemas de Qualidade de Dados

O aumento do volume de geração de dados, dos tipos de dados e da complexidade dos sistemas de informação pode levar ao aparecimento de problemas de qualidade. A mitigação é realizada através da implementação de programas de avaliação e melhorias da qualidade (Orr, 1998). No entanto, para que estes programas sejam implementados, é necessário compreender o conceito da falta de qualidade de dados. A Tabela III apresenta algumas definições de falta de qualidade de dados.

Tabela III - Definição de Falta de Qualidade de Dados

Autor	Definição
(Wang et al., 1993, pg.670)	*(…) data may be of poor quality because it does not reflect real world conditions, or because it is not easily used and understood by the data user.*
(Wand & Wang, 1996, pg. 104)	*A data deficiency is an inconformity between the view of the real-world system that can be inferred from a representing information system and the view that can be obtained by directly observing the real-world system.*
(Strong et al., 1997, pg. 104)	*(…) DQ problem is any difficulty encountered along one or more quality dimensions that renders data completely or largely unfit for use.*
(Kim et al, 2003, pg.82)	*(…) we say that data is dirty if the user or application ends up with a wrong result or is not able to derive a result due to certain inherent problems with the data.*

As deficiências de qualidade podem estar associadas a falhas no desenho do SI, aos processos diários utilizados para registo, armazenamento, tratamento e disponibilização dos dados, como também por influência externa dos utilizadores envolvidos no ecossistema dos dados (Redman, 1998; Strong et al., 1997; Wand & Wang, 1996).

O estudo efetuado por Strong et al. (1997) a 42 projetos de QD permitiu identificar algumas fontes de problemas de qualidade. O estudo concluiu que a origem de falta de qualidade associada às dimensões da categoria intrínseca deve-se à existência de múltiplas fontes de dados e à subjetividade nos processos que produzem os dados. A falta de qualidade, associada aos problemas de acessibilidade, deve-se às dificuldades técnicas na disponibilização de acesso, ao tempo de processamento dos dados que

pode inviabilizar a sua utilização e a problemas de interpretabilidade que dificultam ao utilizador compreender a informação. Quanto à categoria contextual, a falta de qualidade está associada a dados em falta nas tabelas, a dados mal definidos e à dificuldade em realizar agregações de dados.

A taxonomia definida por Oliveira et al. (2005), representada no Anexo II, apresenta de forma detalhada os problemas de qualidade de dados (PQD) que podem ser identificados. A lista foi realizada comparando as investigações efetuadas por Kim et al. (2003), Müller & Freytag (2003) e Rahm & Do (2000).

2.3. Metodologias de Qualidade de Dados

As metodologias permitem identificar as fases a percorrer nos estudos ou projetos. Tendo em mente este objetivo, à medida que o tema de qualidade de dados ganhou importância, tornou-se importante identificar as etapas a percorrer para endereçar os projetos de qualidade (Batini, Cappiello, Francalanci, & Maurino, 2009).

Na literatura existente é possível verificar a existência de várias metodologias de melhoria de QD, como evidenciado pelo estudo realizado por Batini et al (2009). Uma das mais aclamadas pelos autores é a metodologia de gestão total de qualidade de dados (GTQD) (Batini et al., 2009; Wang, 1998).

A metodologia GTQD, introduzida por Wang (1998), nasceu através da analogia do processo de qualidade de fabrico da informação com a qualidade de fabrico de um produto (*Total Quality Management* – TQM).

A metodologia GTQD alia-se à perspetiva de que os dados são produtos de informação (PI) e compara as fases de fabrico do produto às dos produtos de informação, como é possível verificar na Tabela IV. A metodologia promove melhorias contínuas nos processos de qualidade, propondo um ciclo composto por quatro fases denominadas de definição (*Define*), avaliação (*Measure*), análise (*Analyse*), e aplicação de melhorias (*Improve*) (Figura 1).

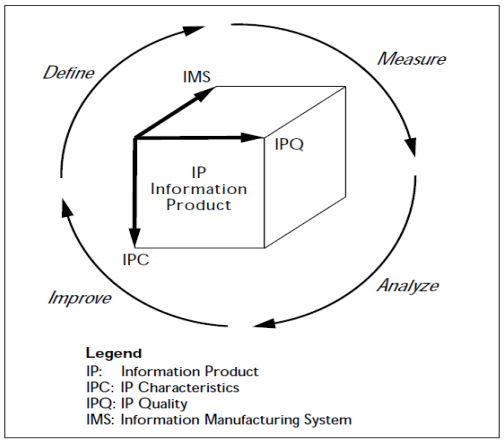
Tabela IV - Produto vs. Produto de Informação

	Product Manufacturing	Information Manufacturing
Input	Raw Materials	Raw Data
Process	Assembly Line	Information System
Output	Physical Products	Information Products

Fonte: Wang, R. (1998), pg. 59

Na primeira fase, definição, a primeira etapa é a análise das características dos dados e dos respetivos requisitos. Este processo permite verificar como estão construídos os processos que produzem os dados, definir os dados que serão analisados e identificar as dimensões de qualidade de dados relevantes no processo de qualidade.

Figura 1 - Ciclo da metodologia GTQD



Fonte: Wang, R. (1998), pg. 60

Na fase seguinte, avaliação, são criadas regras para medir a qualidade dos dados. As regras podem ser de quatro tipos: regras para analisar as dimensões de QD; regras de integridade; regras para validar tabelas de referência e regras de negócio.

Na fase de análise são aplicadas as regras criadas na fase anterior. Este processo permite analisar a origem das deficiências de qualidade e identificar técnicas para as corrigir.

Na última fase são implementadas melhorias de qualidade nas deficiências identificadas na fase anterior.

No projeto em questão foi aplicada a metodologia *SAS Dataflux* que apresenta características semelhantes à GTQD.

Ao contrário da metodologia GTQD a metodologia *SAS Dataflux* é composta por três fases: planeamento, ação e monitorização.

A nível de comparação as atividades desenvolvidas nas fases definição, avaliação e análise da metodologia GTQD são na sua maioria efetuadas na fase planeamento. Na fase ação são efetuadas ações para melhorar a qualidade de dados, tal como, na fase de implementação de melhorias da metodologia GTQD. A maior diferença entre as

duas metodologias é a existência da fase monitorização onde são efetuadas ações para avaliar a eficácia das melhorias aplicadas e onde é possível definir métricas que de forma contínua avaliam a qualidade dos dados.

A metodologia *SAS Dataflux* será apresentada com mais detalhe na fase de projeto.

2.4. Melhoria da qualidade de dados

A atividade de resolução de anomalias de QD é denominada de limpeza de dados (LD) e é composta por três fases: identificação dos erros de QD; definição de medidas a aplicar para melhorar a QD; e aplicação das correções nos dados (Müller & Freytag, 2003; Raman & Hellerstein, 2001).

Na primeira fase auditam-se os dados para identificar as anomalias. Uma das técnicas utilizadas para auditar é o *data profiling* (DP) que examina cada atributo e obtém informação estatística sobre o mesmo, como por exemplo, % de valores a *null*, valor máximo e mínimo, análise de padrões de valores e *outliers* (Naumann, 2014). Os resultados do DP são importantes no processo de melhoria de QD, uma vez que são o suporte para a identificação dos atributos que apresentam problemas de qualidade e as respetivas regras de resolução (Müller & Freytag, 2003).

Na segunda fase identificam-se as medidas a aplicar para melhorar a QD. Nesta fase é necessário considerar que na resolução dos PQD podem ser aplicadas dois tipos de estratégias, nomeadamente, reativa e proativa (Batini et al., 2009; Strong et al., 1997).

A estratégia reativa permite aplicar técnicas de correção pontuais com o objetivo de resolver os problemas que existem na base de dados. A estratégia proativa permite aplicar técnicas de resolução ao nível dos processos que compõem o ciclo de dados. O objetivo desta última é resolver a falta de qualidade corrigindo a origem dos problemas de qualidade de dados (Batini et al., 2009; Huh et al., 1990). Na Tabela V são apresentadas algumas das técnicas utilizadas para melhorar a qualidade de dados.

Tabela V - Técnicas de Qualidade de Dados

Estratégia	Técnica	Definição
Reativa	Aquisição de dados novos	Atualização da tabela com dados novos.
	Normalização	Modificação dos dados considerando valores padrão.
	Entity Resolution	Identificação de registos em uma ou mais tabelas da mesma fonte de dados que representam a mesma entidade, de forma a criar um único registo.
	Integração de Dados	Integração de dados de tabelas provenientes de duas ou mais fontes Existe dois tipos: *Quality driven query* – Integração de dados utilizando fontes com credibilidade; *Instance level conflict resolution* – Identificação e resolução de valores que identificam o mesmo objeto.
	Localização de Erros	Conjunto de regras semânticas para identificar registos, em uma ou mais tabelas, que não respeitem as regras.
	Correção de Erros	Conjunto de regras semânticas para corrigir os tipos de erros detetados em uma ou mais tabelas.
	Data Enrichment	Processo que permite integrar dados de tabelas externas para melhorar informação incompleta
Proativa	Controlo de Processos	Introdução de regras de controlo nos processos que criam os dados.
	Redefinição de Processos	Introdução de novas atividades para produzir dados com melhor qualidade.

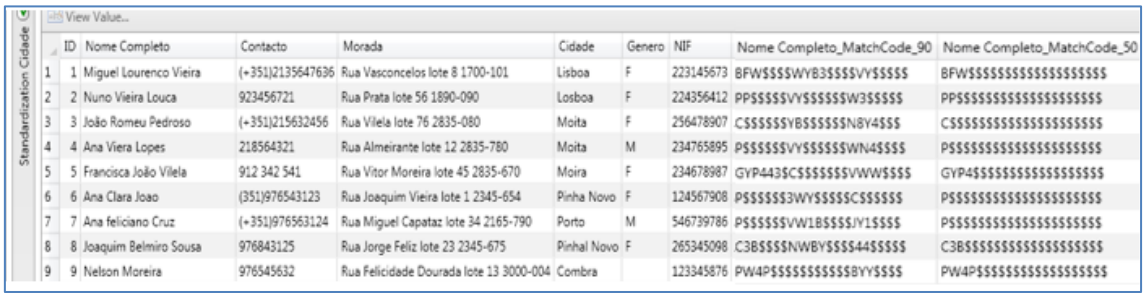
Fonte: Adaptado de Batini et al (2009);Barateiro & Galhardas (2005)

Em detalhe, a normalização contém uma variada lista de atividades que podem ser aplicadas aos dados de forma corrigi-los (Batini e Scannapieco, 2006):

- *Parse*: Atividade para reorganizar a composição dos atributos, separando os valores em *substrings*. Esta atividade é muito utlizada para estruturar atributos de morada ou nome, ou para facilitar a correção de valores incorretos através da utilização de tabelas *lookup*.
- Validação de formato: Atividade que valida e homogeneíza os dados, para que todos os valores do atributo apresentem o mesmo formato.
- Substituição de ortografia alternativa: Substituição de valores abreviados para valores mais completos e compreensíveis.

É de salientar que no processo de enriquecimento de dados (*data enrichement*), considerando o *software dataflux*[1], existe um processo denominado de *match codes* que é chave para a criação das regras que permitem cruzar e enriquecer os dados. *Match codes* são representações codificadas de dados que são semelhantes. A técnica em questão utiliza lógica difusa proprietária do SAS, que codifica os valores dos atributos, considerando um nível de sensibilidade selecionado pelo utilizador, para identificar valores semelhantes nos registos (Queen, 2016). A Figura 2 apresenta um exemplo da codificação realizada pela técnica *match codes*.

Figura 2 - Exemplo da técnica Match Codes



	ID	Nome Completo	Contacto	Morada	Cidade	Genero	NIF	Nome Completo_MatchCode_90	Nome Completo_MatchCode_50
1	1	Miguel Lourenco Vieira	(+351)2135647636	Rua Vasconcelos lote 8 1700-101	Lisboa	F	223145673	BFW$$$$WYB3$$$$VY$$$$	BFW$$$$$$$$$$$$$$$$$$
2	2	Nuno Vieira Louca	923456721	Rua Prata lote 56 1890-090	Losboa	F	224356412	PP$$$$$VY$$$$$$W3$$$$$	PP$$$$$$$$$$$$$$$$$$
3	3	João Romeu Pedroso	(+351)215632456	Rua Vilela lote 76 2835-080	Moita	F	256478907	C$$$$$$YB$$$$$$N8Y4$$$	C$$$$$$$$$$$$$$$$$$
4	4	Ana Viera Lopes	218564321	Rua Almeirante lote 12 2835-780	Moita	M	234765895	P$$$$$$VY$$$$$$WN4$$$$	P$$$$$$$$$$$$$$$$$$
5	5	Francisca João Vilela	912 342 541	Rua Vitor Moreira lote 45 2835-670	Moira	F	234678987	GYP443$C$$$$$$$VWW$$$$	GYP4$$$$$$$$$$$$$$$$
6	6	Ana Clara Joao	(351)976543123	Rua Joaquim Vieira lote 1 2345-654	Pinha Novo	F	124567908	P$$$$$$3WY$$$$$C$$$$$$	P$$$$$$$$$$$$$$$$$$
7	7	Ana feliciano Cruz	(+351)976563124	Rua Miguel Capataz lote 34 2165-790	Porto	M	546739786	P$$$$$$VW1B$$$$JY1$$$$	P$$$$$$$$$$$$$$$$$$
8	8	Joaquim Belmiro Sousa	976843125	Rua Jorge Feliz lote 23 2345-675	Pinhal Novo	F	265345098	C3B$$$$NWBY$$$$44$$$$	C3B$$$$$$$$$$$$$$$$$$
9	9	Nelson Moreira	976545632	Rua Felicidade Dourada lote 13 3000-004	Combra		123345876	PW4P$$$$$$$$$$BYY$$$$	PW4P$$$$$$$$$$$$$$$$

Esta técnica é muito utilizada para identificar dados que estão duplicados (Queen, 2016), já que em vez de verificar se os dados são literalmente iguais permite realizar correspondência se houver pequenas variações (ex. nomes Carolina Machado e Carol Machado).

A utilização das técnicas proactivas permite eliminar a longo prazo as fontes dos PQD, podendo, no entanto, ser dispendiosas a curto prazo. As técnicas reativas são consideradas aconselháveis para a informação estática e mais eficientes a curto prazo, no entanto não eliminam a fonte dos problemas de qualidade (Batini et al, 2009).

Na última fase do processo de limpeza de dados aplicam-se as técnicas definidas para corrigir a qualidade dos mesmos.

Independentemente das técnicas aplicadas, a resolução das deficiências de qualidade é um processo evolutivo de tentativa e erro, isto é, à medida que se vão aplicando regras avalia-se o resultado e promovem-se melhorias (LEE, 2003).

[1] *Sas Dataflux* – O *SAS Dataflux* é um *software* da empresa SAS Institute que contém uma componente de qualidade de dados. Esta componente permite identificar e solucionar problemas de qualidade.

No projeto em questão, face ao seu âmbito, foram aplicadas técnicas reativas para melhorar a qualidade de dados. As regras criadas para melhorar a QD, foram fornecidas à FinanceQ de forma a aplicarem nos processos que iriam gerar dados no novo sistema.

3. Apresentação do Projeto

A empresa FinanceQ, instituição que vende soluções financeiras, no âmbito do projeto de migração da base de dados (IBM AS400) para uma versão mais recente, considerou importante melhorar a qualidade dos dados a migrar para a nova versão. Tendo como objetivo analisar o estado atual dos dados e implementar medidas de qualidade a FinanceQ requisitou os serviços da SIGQ, empresa que fornece soluções no setor dos sistemas de informação. O projeto apresentado será focado na fase de intervenção sobre a qualidade de dados.

No âmbito do projeto foram propostos os seguintes objetivos:

- Análise à qualidade dados: análise de anomalias utilizando técnicas de *data profiling*, aplicação de regras de negócio e de validação de tabelas de referência;
- Normalização de dados: limpeza de dados descontextualizados, correção de erros ortográficos e normalização da forma; e
- Enriquecimento de moradas: normalização das moradas segundo formato definido e enriquecimento de códigos postais de 7 dígitos utilizando informação pública disponibilizada pelos CTT.

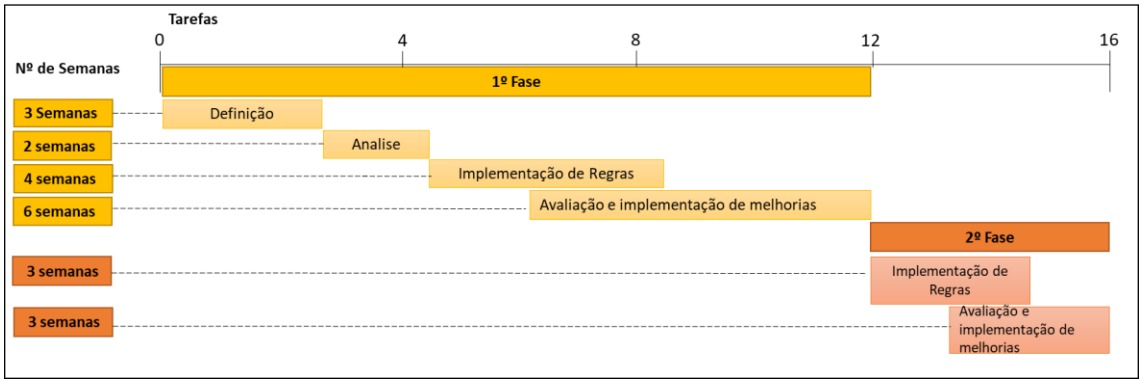
De forma a cumprir os objetivos propostos o projeto foi dividido em duas fases:

1º Fase: Identificação de anomalias e aplicação de melhorias; e

2º Fase: Enriquecimento dos dados relativos a moradas.

O projeto teve uma duração de quatro meses. Nos primeiros três meses foram efetuadas as tarefas relativas à primeira fase e no último mês efetuaram-se tarefas relativas à segunda fase. Este planeamento está espelhado na Figura 3.

Figura 3 - Calendário do planeamento de tarefas

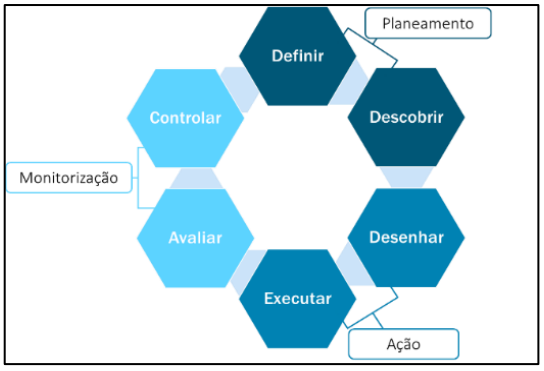


Para identificar e solucionar as anomalias de qualidade o cliente adquiriu o *software SAS Dataflux*.

Este *software* aplica a metodologia *SAS Dataflux* (Figura 4), composta por três fases, que por sua vez, se subdividem em duas:

- Planeamento - Definição do âmbito, fontes de dados, exploração dos dados e identificação de ações a realizar.
- Ação – Definição e Implementação de ações de qualidade.
- Monitorização – Avaliação e retificação da implementação. Definição de métricas de monitorização contínua.

Figura 4 - Metodologia SAS Data



Como é possível verificar na Figura 4, o planeamento é dividido em duas subfases, definir e descobrir respetivamente. A fase ação é composta pelas subfases desenhar e executar. Por fim, a fase monitorização é composta pelas subfases avaliar e controlar.

3.1. Ciclo de Melhoria de Qualidade de Dados

A melhoria de qualidade de dados (MQD) realizou-se considerando as fases da metodologia *SAS Dataflux*. Na primeira fase, planeamento, identificaram-se os dados a

analisar e os problemas de qualidade. Na fase seguinte, ação, procedeu-se ao desenho e implementação das medidas corretivas. Na última fase, monitorização, procedeu-se à avaliação comparando os resultados iniciais do *data profiling* com a avaliação realizada após aplicação das MQD, e considerando as avaliações realizadas pelo utilizador final.

Apesar da metodologia aplicada conter apenas três fases é em muito semelhante à metodologia GTQD. As atividades desenvolvidas nas primeiras três fases da metodologia GTQD estão englobadas na fase planeamento, e a última fase, implementação de melhorias, está englobada na fase ação. A diferença da metodologia GTQD prende-se com a introdução da fase controlo, em que se realiza a avaliação das melhorias aplicadas. É de notar que na metodologia *SAS Dataflux* o foco não é a identificação das dimensões de qualidade a melhorar, mas sim os valores dos atributos que devem ser melhorados.

As alíneas 3.1.1 a 3.1.5 espelham os procedimentos que foram efetuados durante o projeto nas três fases da metodologia *SAS Dataflux*.

3.1.1. Definição do âmbito

A primeira atividade do processo de avaliação de qualidade na metodologia aplicada é a identificação dos dados a analisar. No âmbito do projeto foram analisadas duas tabelas de dados que caracterizam as entidades: cliente final e pontos de venda. A Tabela VI apresenta a caracterização das respetivas tabelas.

Tabela VI - Caracterização Tabelas Clientes e Ponto de Venda

TABELA	Nº COLUNAS	Nº TUPLOS
CLIENTES	84	2381389
PONTO VENDA	94	24477

Considerando a taxonomia de Oliveira et al (2005) apresentada na secção 2.2 e em detalhe no Anexo II, aplicou-se a técnica *data profiling* para identificar as anomalias nos dados. O Anexo I contém o detalhe das métricas analisadas pelo DP do *software SAS Dataflux.*

A melhoria de qualidade incide principalmente nos dados dos atributos que caracterizam as entidades e que a FinanceQ necessita para contactar corretamente o cliente e/ou parceiro. O Anexo III apresenta a lista de atributos em detalhe que o negócio considerou como caracterizadores na tabela de clientes. A Tabela VII contém os tipos de atributos, relativos a tabela de clientes, que serão analisados

Tabela VII - Tipos de Atributos Caracterizadores da Tabela de Clientes

TIPO DE ATRIBUTO	QUANTIDADE DE ATRIBUTOS
DATA	5
DOCUMENTO	6
GÊNERO	1
TÍTULO DE CLIENTE	1
NOME CLIENTE	1
TIPO CLIENTE	1
DESIGNAÇÃO TRABALHO	1
TELEFONE	5
EMAIL	2
MORADA	6

O Anexo IV apresenta a lista de atributos em detalhe que o negócio considerou como caracterizadores da tabela de pontos de venda. A Tabela VIII contém os tipos de atributos, relativos a tabela de pontos de venda, que serão analisados.

Tabela VIII - Tipos de Atributos Caracterizadores da Tabela de Pontos de Venda

ATRIBUTO	QUANTIDADE DE ATRIBUTOS
NOME	1
DESIGNAÇÃO COMERCIAL	1
GÊNERO	1
DOCUMENTO	11
EMAIL	1
SITE	1
DATA	6

Para além dos atributos caracterizadores das entidades, foram aplicadas regras para validar a integridade referencial. Esta validação incidiu nos atributos listados na Tabela IX.

Tabela IX - Atributos para Validação de Integridade Referencial

TABELA	ATRIBUTO	DESCRIÇÃO
CLIENTE	CLICAE	CAE Cliente
CLIENTE	ESTCIV	Estado Civil
CLIENTE	TIPNAC	Tipo de Nacionalidade
CLIENTE	TIPPROF	Código Profissão
CLIENTE	CONTTRA	Tipo de Contrato Trabalho
CLIENTE	AGCAECLI	Agrupamento CAE Cliente
CLIENTE	AGCAEEMP	Agrupamento CAE Empresa
CLIENTE	OUTNACIONA	Outra Nacionalidade
CLIENTE	NATURALID	Naturalidade
CLIENTE	CODENTEMI	Cód. Entidade Emitente
CLIENTE	AUFRENDEST	Aufere Rendimento Desde
PONTO VENDA	VNDTOPR	Tipo de Operador
PONTO VENDA	VNDTBEM	Tipo de bem
PONTO VENDA	VNDTBEM2	Tipo de bem secundário
PONTO VENDA	VNDMARC	Marca
PONTO VENDA	VNDMENC	Tipo de Motivo de Encerramento

3.1.2. Diagnóstico de PQD

Após a identificação dos atributos considerados como importantes na análise de PQD, a etapa seguinte da metodologia é a utilização do DP para identificar os PQD.

Os atributos do tipo telefone contêm informação de contacto com o cliente e estão organizados consoante o tipo de contacto. A análise do DP ao atributo TELCASA (telefone de casa) apurou que em 39% dos tuplos o atributo apresentava o valor 0, utilizado quando não há conhecimento do valor. Para além de valor em falta, também foram detetados vários valores sem significado, como também padrões inadequados para números de telefone, por exemplo ‘99’ ou ‘9999999999’. Os mesmos tipos de anomalias foram detetados nos restantes atributos do tipo telefone.

Os atributos do tipo *email* contêm informação do endereço eletrónico que pode ser utilizado para contactar o cliente e subdividem-se em email pessoal e de trabalho. A análise do DP ao atributo EMAIL (email) identificou que 83,24% dos tuplos

apresentavam valores em falta. Para além de valores em falta, detetaram-se valores que não cumprem a estrutura esperada do atributo, e valores com erro ortográfico no domínio do *email*. A análise DP ao atributo VNDINTER (*site*) identificou anomalias semelhantes ao atributo do tipo email, no qual 93,50% dos tuplos apresentavam valores em falta. Foram detetados, à semelhança, erros de ortografia nos domínios e valores sem significado.

Os atributos do tipo documento contêm informação sobre os documentos de identificação das entidades. Neste tipo de atributos a análise do DP foi utilizada para verificar se existem valores sem significado e a existência de registos duplicados. Pela natureza e composição da tabela de pontos de venda, não é possível verificar a existência de registos duplicados através da análise do atributo VNDNFC (NIF coletivo). O cliente assegurou, no entanto, que ambas as tabelas não apresentam registos duplicados. Na tabela de clientes a métrica unicidade do atributo NUMCONT (NIF) apresentou o valor de 100% o que significa que 100% dos registos são únicos. No atributo NUMCARTA (carta de condução) foram identificados valores sem significado, o que também foi possível verificar nos atributos que armazenam informação do NIB.

Nos atributos do tipo género, que guardam informação do género da entidade, a análise DP permitiu apurar que no atributo CLISEX (género) 1,55 % dos tuplos apresentam valores em falta e 0,0002% dos tuplos apresentam valores sem sentido. No atributo VNDGEN (género) apurou que 4,27% dos tuplos apresentam valores em falta.

A análise dos atributos tipo cliente, cujo objetivo é identificar o título que se utiliza para se dirigir ao cliente, permitiu identificar a existência de sinónimos, 1,36% dos tuplos com valores em falta e vários tuplos que apresentam valores sem sentido. A análise DP ao atributo TIPCLI (tipo de cliente), que caracteriza se o cliente é particular ou empresa, não evidenciou problemas de qualidade. A análise DP apurou que 1,83% dos tuplos representam empresas e que 98,17% representam particulares.

A análise ao atributo que identifica a profissão do cliente, TITTRA (profissão), apurou a existência de um grande número de tuplos com valor sem significado, erros

ortográficos, existência de sinónimos e que 40,69% dos tuplos apresentam valores em falta.

Nos atributos do tipo data o objetivo é verificar se apresentam a sintaxe definida para o atributo e identificar se existem valores sem significado. Em detalhe a análise DP no atributo DATANI (data de nascimento) permitiu verificar que contém padrões inadequados para data, por exemplo '9999' quando se espera obter '99999999', e que contém datas superior ao dia atual, no entanto não existem valores em falta. Na análise DP do atributo EMPDAT (data de contratação), para além das anomalias evidenciadas na análise ao atributo DATANI (data de nascimento), foi possível identificar que 19,69% do tuplos contêm valor 0, que é utilizado quando não há conhecimento do valor. Os mesmos tipos de anomalias foram evidenciados nos restantes atributos do tipo data.

Nos atributos do tipo morada a finalidade é registar informação para que seja possível enviar correspondências ao cliente. A análise DP ao atributo MORADA (morada) apurou a existência de valores sem significado e erros ortográficos. O atributo NUMRUA (número da rua) apresenta valores sem significado. No atributo LOCALID (localidade) a análise DP apurou a presença de erros ortográficos, valores sem significado e 40% dos registos com este atributo a *null*.

A análise aos atributos do tipo nome, que representam informação sobre o nome do cliente, permitiu apurar problemas tais como, presença de erros ortográficos, valores fora do domínio, valores sem sentido e registos com valor em falta. Em detalhe, no atributo VNDDCOM (designação comercial), a análise DP identificou a existência de sinónimos na forma jurídica.

3.1.3. Desenho de Medidas de Correção

A etapa seguinte à identificação das anomalias, com o auxílio da ferramenta *data profiling*, é a definição das regras que serão implementadas para corrigir os dados.

As regras de validação definidas para corrigir os PQD agrupam-se em quatro categorias:

- Integridade existencial/conteúdo – Mede a caracterização do conteúdo do campo;
- Regras de negócio – Valida a qualidade de dados quanto a regras próprias do negócio.
- Integridade referencial – Validação entre valores de tabelas de referência e os respetivos valores nas tabelas que os referenciam.

No geral, nos atributos do tipo data, foram aplicadas regras para garantir que todas as datas apresentavam mesma sintaxe, isto é, 'AAAAMMDD'. Os valores que forem menores de oito dígitos serão mantidos em branco. Para o atributo DATANI (data de nascimento) delineou-se uma regra para garantir que apenas se mantêm datas de nascimento em que o cliente tenha pelo menos dezoito anos. Esta regra foi implementada para garantir que a data representada é a que tem mais probabilidade de corresponder a situação real.

Nos atributos do tipo telefone definiram-se regras para garantir que todos os valores de clientes com residência em Portugal apresentam o mesmo formato '999999999'. Todos os valores sem significado ou que começam por 1 serão representados por '0'. Para além das regras anteriores, o cliente apresentou a necessidade de reorganizar os números de telefone:

- TELCASA (telefone de casa) – Para números que começam por 9, passar para TELMOV (telemóvel) se estiver '0'. Caso contrário, passar para o campo TELMTRB (telemóvel de trabalho) se estiver '0'. Caso contrário mantém valor no atributo TELCASA (telefone de casa).
- TELMOV (telemóvel) – Para números que começam por números diferentes de 9 passar para TELCASA (telefone de casa) se estiver '0'. Caso contrário, passar para TELTRBO (telefone de trabalho) se estiver '0'. Caso contrário manter valor no atributo TELMOV (telemóvel).
- TELTRBO (telefone de trabalho) – Para números que começam por 9, passar para TELMTRB (telemóvel de trabalho) se estiver '0'. Caso contrário, passar para TELMOV (telemóvel) se estiver '0'. Caso contrário manter valor no atributo TELTRBO (telefone de trabalho).

- TELMTRB (telemóvel de trabalho) – Para números que começam por números diferentes de 9 passar para TELTRBO (telefone de trabalho) se estiver '0'. Caso contrário, passar para TELCASA (telefone de casa) se estiver '0'. Caso contrário manter valor no atributo TELMTRB (telemóvel de trabalho).
- TELFAX (FAX) - Para números que começam por 9, passar para TELMOV (telemóvel) se estiver '0'. Caso contrário, passar para o campo TELMTRB (telemóvel de trabalho) se estiver '0'. Caso contrário mantém valor no atributo TELFAX (FAX).

Nos atributos EMAIL (email), EMAILT (email de trabalho) e VNDINTER (site) foram delineadas medidas para validar o formato e regras de normalização para corrigir os erros ortográficos nos domínios do email e do site. Os valores que não cumprem o formato esperado passam a estar a *null*.

Nos atributos do tipo género foram definidas regras para verificar se o valor atribuído é o correto. No atributo CLISEX (género) a regra aplicada considera que o atributo deve estar preenchido se a entidade for de nacionalidade portuguesa e se o TIPCLI (tipo de cliente) for particular. De seguida valida o género consoante o primeiro nome da entidade, disponível no atributo CLINOM (nome). Nos casos em que a entidade é empresa o atributo deve apresentar valor a *null*. No atributo VNDGEN (género) a regra aplicada para validar os géneros considera se a entidade é particular ou empresa através do NIF e se for particular valida o género consoante o primeiro nome da entidade, disponível no atributo VNDNOME (nome). Nos registos em que a entidade é empresa o atributo é preenchido com o valor 'E'. A finalidade das medidas a aplicar é garantir que os dados representam exatamente a situação real e melhorar a completude dos registos, preenchendo se possível os campos que estão a *null*

No atributo TIPCLI (tipo de cliente) a validação realizou-se considerando o NIF, isto é, se o NIF começar por [1,2] então o tipo de cliente é particular, se começar por [5,6,8,9] é empresa. O atributo de validação NIF é o NUMCONT. Como regra de negócio o cliente adicionou a necessidade de validar que se o TIPCLI (tipo de cliente) for particular o género deve estar preenchido e o atributo CLICAE (CAE) não deve estar

preenchido; se o TIPCLI (tipo de cliente) for empresa o género e número de BI não devem ser preenchidos.

No atributo CLITIT (título do cliente) foram definidas medidas para retirar os valores sem significado e medidas de normalização para que as abreviaturas apresentem a mesma estrutura. Na análise DP foi detetado no atributo CLINOM (nome) valores que continham informação sobre o título do cliente. O objetivo será passar a parte do valor referente a informação do título do cliente para o atributo CLITIT (título do cliente) caso este esteja a *null*. A regra implementada para garantir que o título está correto valida consoante o género se os valores 'SR', ou 'SRA' estão corretamente designados, e para os registos em que o CLITIT (título do cliente) está vazio e tipo de cliente é particular regista os valores consoante o género.

No atributo CLINOM (nome), VNDNOME (nome) e VNDDCOM (designação comercial) definiram-se técnicas de normalização para corrigir erros ortográficos e retirar os valores sem significado. Nos atributos CLINOM (nome) e VNDNOME (nome) pretende-se retirar os acentos e os elementos de ligação. No atributo VNDDCOM (designação comercial) a parte do valor que corresponde à forma jurídica será normalizada de forma a eliminar a existência de sinónimos. A própria estrutura do campo será formatada de forma a haver mais consistência entre os valores do atributo.

No atributo TITRA (profissão) foram aplicadas medidas de normalização para retirar os valores sem significado, eliminar a existência de sinónimos e melhorar a consistência entre os valores do atributo. A finalidade é disponibilizar uma lista uniforme para corrigir os valores que existem no atributo TITTRA (profissão).

Para os atributos NUMBI (Bilhete de identidade), NUMCONT (NIF), NPCNUM (NIF Coletivo) e VNDNFC (NIF) aplicou-se o algoritmo Módulo 11[2] que valida o dígito de controlo. Para os atributos NIB aplicou-se o algoritmo IBAN[3] que válida os dígitos de controlo do NIB. Estas validações apenas foram realizadas para os NIFs e NIBs portugueses.

[2] Este algoritmo é utilizado para validar o NIF, verificando o dígito de controlo.

[3] Este algoritmo valida se o IBAN é válido verificando os dois últimos dígitos de controlo. O procedimento de cálculo está disponível no site do Banco de Portugal: https://www.bportugal.pt/sites/default/files/anexos/documentos-relacionados/international_bank_account_number_pt.pdf

Para o atributo NUMCARTA (carta de condução) por não existir algoritmo validação foram aplicadas medidas de limpeza nos valores identificados como inadequados.

Para os atributos em que se pretende verificar a integridade referencial, aplicaram-se regras de validação para garantir que os valores apresentados existem na respetiva tabela de referência.

Nos atributos de morada o objetivo é enriquecer os atributos do código postal (LOCALID e CODPOS). Este trabalho foi efetuado utilizando a informação de morada do cliente alocada nos atributos NUMRUA, MORADA, PORTA e ANDAR e comparando com a base de dados de moradas dos CTT.

A base de dados de comparação foi obtida no site institucional dos CTT que disponibiliza gratuitamente esta informação sob forma de ficheiros[4]. A lista de ficheiros utilizada é composta por quatro ficheiros: lista de distritos; lista de concelhos; lista de freguesias; e lista de códigos postal. O objetivo será criar regras para comparar as moradas dos clientes da FinanceQ com as moradas dos CTT e enriquecer os atributos de informação de código postal.

O processo de enriquecimento das moradas é composto por três etapas. Na primeira etapa foram efetuadas ações de preparação das tabelas dos CTT para efetuar o cruzamento de dados entre as tabelas. Na segunda etapa foram realizadas ações de normalização (separação, correção de erros ortográficos e remoção de acentuação) nos dados da FinanceQ segundo o formato disponibilizado pelos CTT (ex. Tabela X). Na última etapa as regras definidas foram aplicadas e procedeu-se ao preenchimento dos atributos de código postal.

[4] Site de obtenção de ficheiros:
https://www.ctt.pt/feapl_2/app/restricted/postalCodeSearch/postalCodeDownloadFiles.jspx

Tabela X - Exemplo Normalização considerando o formato dos CTT

Exemplo Morada	Layout CTT	Exemplo Morada Separada
Avenida João de Deus nº 40 2840-010 Casal do Marco	Tipo Rua	Avenida
	Morada	Joao de Deus
	Porta	40
	Código Postal (2 dígitos)	28
	Código Postal (4 dígitos)	2840
	Código Postal (3 dígitos)	010
	Localidade	Casal do Marco

3.1.4. Implementação de Medidas de Correção

A etapa seguinte ao desenho das regras é a fase de implementação. As regras desenhadas foram implementadas recorrendo à ferramenta *SAS Dataflux*.

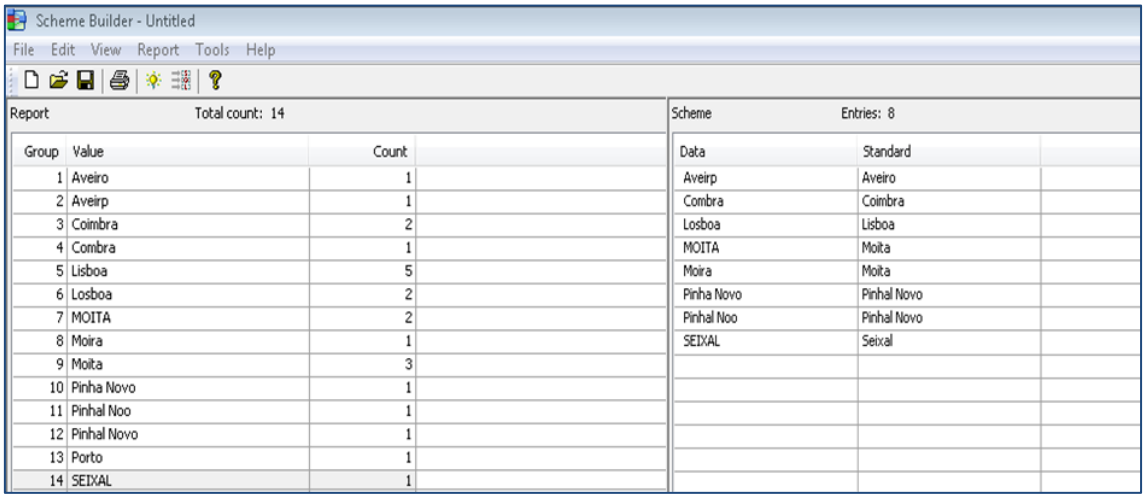
Nos atributos de data e de telefone aplicaram-se as regras de transformação dos formatos, de validação de tamanho e reorganização de valores utilizando a linguagem EEL (*Expression Language*) (Anexo V).

Para os atributos de género executaram-se as tarefas previstas para validar os géneros designados e enriquecer os que apresentavam erradamente o valor a *null*. A tarefa de designação de género foi composta por duas fases: (i) *parse* do atributo CLINOM (nome) em primeiro nome, nomes do meio e último nome (ii) designação do género consoante o primeiro nome. A análise de género apenas foi possível porque o *software* utilizado já continha um largo dicionário de nomes portugueses e regras que designam a probabilidade de o nome ser do género feminino ou masculino. No atributo CLITIT (título do cliente) foram efetuadas tarefas semelhantes.

Para resolver os erros ortográficos e retirar os valores sem significado, uma das técnicas de normalização que o *dataflux* disponibiliza é a ferramenta *scheme*. Esta ferramenta é composta pela lista de valores dos atributos e permite ao utilizador designar o valor de substituição. Esta lista é realizada manualmente e aplicada na fase de implementação. Como é possível verificar na Figura 5, o *scheme* contém na parte

esquerda da janela a lista de valores do atributo em análise, e na parte direita os valores a corrigir e as respetivas retificações.

Figura 5 - Exemplo da ferramenta *Scheme*



Group	Value	Count
1	Aveiro	1
2	Aveirp	1
3	Coimbra	2
4	Combra	1
5	Lisboa	5
6	Losboa	2
7	MOITA	2
8	Moira	1
9	Moita	3
10	Pinha Novo	1
11	Pinhal Noo	1
12	Pinhal Novo	1
13	Porto	1
14	SEIXAL	1

Data	Standard
Aveirp	Aveiro
Combra	Coimbra
Losboa	Lisboa
MOITA	Moita
Moira	Moita
Pinha Novo	Pinhal Novo
Pinhal Noo	Pinhal Novo
SEIXAL	Seixal

Para a validação dos documentos as regras dos algoritmos de validação do *check digit* foram recriadas utilizando a linguagem EEL e aplicadas para verificar a validade dos NIFs e dos NIBs. Apenas aplicável em NIFs e NIBs portugueses. Os valores que não são válidos foram disponibilizados ao cliente, para que possa averiguar.

Nos atributos CLINOM (nome), VNDNOME (nome), VNDCOM (designação comercial) aplicaram tarefas de normalização compostas por três fases: (i) parse dos valores do atributo; (ii) aplicação de *schemes* para resolver os erros ortográficos, retirar valores sem significado e eliminar a existência de sinónimos; (iii) aplicação de expressões regulares para retirar os elementos de ligação e acentos ortográficos.

Para as regras de integridade referencial utilizou-se a funcionalidade *business rules (BR)* do *dataflux* que permite criar regras que averiguam se os valores que existem num determinado atributo existem na lista de valores da tabela de referência. Os valores que não constavam na lista de valores da tabela de referência foram eliminados. As BRs foram aplicadas em cada atributo que valida com tabelas de referência.

Quanto ao processo de enriquecimento, na primeira etapa as tabelas dos CTT foram estruturadas para que fosse possível cruzar os dados com as tabelas da FinanceQ.

Na segunda etapa os atributos foram separados considerando o formato dos CTT e os erros ortográficos e os valores sem significado foram retirados, ou seja, os dados foram normalizados. Para cruzar os dados, ainda nesta etapa, criaram-se atributos de auxílio utilizando a técnica *match codes.* Esta técnica permite criar atributos com diferentes níveis de sensibilidade. Os atributos criados são a chave para a criação dos critérios de cruzamento. A Tabela XI contém exemplos dos *match codes* criados.

Tabela XI - Exemplo Atributos *Match Codes* Calculados

Ex. Match Codes Calculados	
Arruamento_80	Arruamento (Tipo Rua, morada e porta) a 80%
Arruamento_50	Arruamento (Tipo Rua, morada e porta) a 50%
CP2	Primeiros dois dígitos do código postal
CP3	Três últimos dígitos do código postal
Localidade_50	Localidade a 50%
Localidade_80	Localidade a 80%

Na última fase as regras de cruzamento foram aplicadas. A Tabela XII apresenta exemplos de regras criadas utilizando os *match codes* para enriquecer os códigos postais e as localidades.

Tabela XII - Exemplo Regras de Cruzamento

Ex. Critérios de Cruzamento	
Tabela Clientes	Tabela CTT
CP2 + Localidade_50 + Arruamento_50 =	CP2 + Localidade_50 + Arruamento_50
CP2 + Arruamento_80 =	CP2 + Arruamento_80
CP2 + Localidade_80 =	CP2 + Localidade_80

3.1.5. Avaliação Medidas Corretivas

A implementação das regras nos atributos de género permitiu corrigir alguns registos com género incorreto e diminuir a quantidade de tuplos com este atributo a *null* por desconhecimento do valor. Como é possível verificar na Tabela XIII, no atributo CLISEX (género) foi possível verificar que cerca de 0,85% dos tuplos apresentavam erradamente a designação de 'M' e que em vez de 1,55% de valores a vazio por desconhecimento, passou a haver apenas 0,00189% de tuplos.

Tabela XIII - Resultados da MQD nos atributos de género

Atributo	Antes MQD					Depois MQD				
	Nº Tuplos	%F	%M	%E	%Vazio	Nº Tuplos	%F	%M	%E	%Vazio
CLISEX	2381389	48,15	50,3	-	1,550	2381389	48,71	49,45	1,84	0,00189
VNDGEN	24477	4,71	26,13	64,89	4,27	24477	4,27	26,10	64,93	4,27

Nos atributos de *email* e site os *schemes* criados permitiram corrigir os domínios que apresentavam erros e eliminar os valores sem significado. A Tabela XIV apresenta exemplos de *emails* e sites e os tipos de transformação que sofreram para ficarem corretos.

Tabela XIV - Exemplos correções efetuadas nos *emails* e sites

Ex. valores do atributo antes do MQD	Valor dos atributos após MQD
xxxx.xxxx@gmaiil.com	xxxx.xxxx@gmail.com
xxxx.xxxx@gmail.comm	xxxx.xxxx@gmail.com
xxxxx.xxx@yahooo.pt	xxxxx.xxx@yahoo.pt
xxxxx.xxx@hotmaail.comm	xxxxx.xxx@hotmail.com
www.xxxxx.comm	www.xxxxx.com
www.xxxx.ptt	www.xxxx.pt

Para além da transformação efetuada para corrigir erros ortográficos, também foram aplicadas medidas para eliminar valores sem significado. A Tabela XV apresenta o resultado da tarefa de limpeza.

Tabela XV - Resultado da eliminação de valores sem significado nos *emails* e *site*

	Nº tuplos	Nº tuplos preenchidos	Nº tuplos com valores sem significado
EMAIL	2381389	399028	1909
EMAILT	2381389	1302	51
VNDINTER	24477	1592	23

A implementação das regras definidas nos atributos telefone permitiu reorganizar os números, tornando os atributos TELCASA (telefone de casa) e TELMOV (telemóvel) mais completos, eliminar os valores sem significado e garantir que todos os atributos do tipo telefone apresentam o mesmo formato. A Tabela XVI apresenta exemplos das melhorias efetuadas. Note-se que, por questões de privacidade de informação, os dados são fictícios.

Tabela XVI - Exemplo de correções nos atributos de telefone

Tipo de correção	Antes MQD		Após MQD	
	TELCASA	TELMOV	TELCASA	TELMOV
Reorganização valores	91XXXXXXX		0	91XXXXXXX
		21XXXXXXX	21XXXXXXX	0
	91XXXXXXX	96XXXXXXX	91XXXXXXX	96XXXXXXX
Eliminação valores sem significado	210000000	91XXXXXXX	0	91XXXXXXX
	91XXXXXXX	910000000	0	91XXXXXXX

As regras implementadas nas datas, utilizando código EEL, permitiram validar que todas as datas apresentam a mesma sintaxe. Os valores que não respeitavam o formato definido por terem menos dígitos passaram a ter o valor a *null*.

No atributo CLITIT (título de cliente), uma vez designados corretamente os géneros, foi possível identificar e corrigir os tuplos em que estava designado 'SR' mas a pessoa é do género feminino, o mesmo aconteceu ao contrário para o valor 'SRA'.

Após implementação das regras no atributo TIPCLI (tipo de cliente), que contém informação sobre o tipo de cliente, foi possível identificar e corrigir registos que estavam erradamente assinalados como particular ou empresa. Analisando os resultados apresentados na Tabela XVII é possível verificar que cerca de 0,81% dos tuplos estavam erradamente assinalados como clientes particulares.

Tabela XVII - Resultados da MQD no atributo TIPCLI

	Antes da MQD		Após MQD	
	Nº Tuplos	%Tuplos Preenchidos	Nº Tuplos	%Tuplos Preenchidos
PARTICULAR	2381389	98,17%	2381389	97,36%
EMPRESA	2381389	1,83%	2381389	2,64%

No atributo TITTRA (profissão) utilizou-se o *scheme* para eliminar os valores sem significado e a existência de sinónimos. A lista entregue apresentou no total cerca de 721 valores possíveis distintos.

Nos atributos CLINOM (nome), VNDNOME (nome), VNDDCOM (designação comercial) foram aplicadas técnicas de normalização para retirar erros ortográficos, valores sem significado e pontuação errónea. Em particular, no atributo CLINOM (nome), identificaram-se tuplos com a informação do título de cliente que foram retirados e

colocados no atributo CLITIT (título de cliente) se o mesmo estivesse a *null*; no atributo VNDDCOM (designação comercial) aplicaram-se medidas para estruturar o campo e retificar as formas jurídicas. A Tabela XVIII apresenta exemplos das correções efetuadas nos atributos CLINOM (nome), VNDNOME (nome) e VNDDCOM (designação comercial). Note-se que, por questões de privacidade de informação, os dados são fictícios.

Tabela XVIII - Exemplos correções efetuadas nos atributos CLINOM, VNDNOME e VNDDCOM

Tipo de Anomalia	Antes da MQD	Após MQD
PQD: Valor para além do pretendido	DR MIGUEL LOPES	MIGUEL LOPES
	SR JOSE LOPES	JOSE LOPES
	SRA MARA LOPES	MARA LOPES
Formato constante	joao lopes	JOAO LOPES
	XXXXX SA	XXXXX, SA
PQD: Violação Sintaxe	XXXXX LD	XXXXX, LDA
	XXXXX S.A.	XXXXX, SA
	XXXXX LDA.	XXXXX, LDA
	JOAO M.V. LOPES	JOAO M V LOPES
PQD: Erro ortográfico	JOOO LOPES	JOAO LOPES

Quanto às tabelas de referência, as regras implementadas permitiram verificar que os atributos de nacionalidade e CAE apresentavam incoerências quando validados com as respetivas tabelas de referência. Após análise verificou-se que os atributos continham valores que estavam desatualizados. Na correção foram criados *schemes* para corrigir os valores utilizando as regras fornecidas pelo utilizador. Todos os valores que não apresentaram correspondência foram eliminados. A Tabela XIX apresenta as regras utilizadas para corrigir as nacionalidades.

Tabela XIX - Regras para correção nos atributos de nacionalidade

Atributo	Antes MQD	Após MQD
OUTNACIONA (nacionalidade) NATURALID (naturalidade)	A	
	B	
	C	
	D	
	Z	
	E	ES
	F	FR
	P	PT

Quanto aos documentos NIF e bilhete de identidade a validação identificou uma pequena quantidade de valores inválidos. Sendo que nestes casos não foram eliminados, mas sim identificados e fornecidos à FinanceQ para que corrija. Nos atributos que contêm NIBs identificou-se uma pequena quantidade de NIBs inválidos. No entanto a FinanceQ declarou que não pretende que os valores sejam eliminados, uma vez que, são valores default utilizados noutros processos. Quanto ao atributo NUMCARTA (carta de condução), apenas se aplicaram medidas de limpeza utilizando o *scheme* nos valores sem significado.

O processo realizado nas moradas permitiu enriquecer os códigos postais e as localidades. Este processo que teve de duração um mês apresentou mais do que uma iteração pela dificuldade em criar e refinar regras que se adequam a todos os casos que se pretendem enriquecer. No final do processo 80% dos registos (tabela de clientes – 1.905.111 registos) apresentaram os códigos postais e localidade correspondente à base de dados dos CTT. Uma vez que a percentagem de sucesso se enquadrava com a margem pedida pela FinanceQ o processo de enriquecimento foi dado como concluído.

4. Reflexão e Aprendizagem

Este projeto permitiu conhecer os problemas de QD que podem existir nas bases de dados, tal como as técnicas e as validações que podem ser aplicadas para resolver os respetivos problemas. A nível técnico a utilização da ferramenta *SAS Dataflux* possibilitou perceber como as diferentes componentes do *software* podem ser utilizadas para identificar e corrigir problemas de QD. É de salientar que apesar das técnicas implementadas terem sido reativas, a literatura permitiu compreender a importância da reestruturação ao nível dos processos que geram os dados.

Quanto as medidas aplicadas, através do projeto, a FinanceQ compreendeu a importância de identificar corretamente os atributos que devem ser validados por uma lista de valores, como o atributo TITRA (profissão).

Nos números de telefone verificou-se que é importante implementar regras para validar a quantidade de dígitos que são inseridos no campo, tal como os atributos que devem ser de preenchimento prioritário. Sendo que os números de telefone de casa e telemóvel foram identificados como prioritários.

No que se refere às datas a FinanceQ entendeu que é importante criar regras de validação nas datas de nascimento. Nas restantes datas verificou que é importante verificar o padrão que se segue, tal como, a quantidade de dígitos.

Quanto aos atributos de nome as técnicas aplicadas permitiram corrigir os erros ortográficos. Para estes tipos de atributos não existem validações que podem ser implementadas, pelo que deve haver cuidado na inserção dos valores. Quanto aos atributos que apresentaram valores fora do domínio como o NOME (nome do cliente) identificou-se que a razão não é a inexistência do atributo para conter a informação, mas sim, pelo descuido do utilizador aquando da inserção.

Os campos de *email* também apresentaram preocupações para a FinanceQ. As correções aplicadas permitiram limpar os valores sem significado e corrigir os subdomínios (“gmail”) e os domínios (“pt”).

Quanto às moradas, o processo efetuado permitiu destacar a importância dos atributos de código postal e localidade não serem de preenchimento livre.

No que se refere às tabelas de referência a FinanceQ compreendeu a importância de atualizar os valores com mais frequência, uma vez que, os erros despoletados neste âmbito deveram-se a valores desatualizados.

Algumas das medidas de validação sugeridas à FinanceQ foram tornar o atributo CLITIT (título do cliente) obrigatório seguindo uma lista de valores, uma vez que é a chave para uma correta saudação ao cliente, e implementar medidas para validar o padrão de valores dos atributos email, telefones e datas. Quanto ao género não foram sugeridas medidas de validação, uma vez que para validar seria necessário identificar o género considerando o primeiro nome do cliente.

Utilizando as técnicas de *data profiling* e a taxonomia de PQD de Oliveira et al. (2005) foi possível identificar os problemas de qualidade que as tabelas caracterizadoras de

clientes e pontos de venda da FinanceQ apresentavam. Os resultados do DP serviram de base para a criação das regras. Considerando as dimensões de qualidade identificadas por Wang e Strong (1996) e analisando as regras aplicadas, as correções tiveram como objetivo melhorar a exatidão e objetividade dos valores (ex: correção dos erros ortográficos); tornar os dados mais completos (ex: enriquecimento com a análise de género); e melhorar a consistência dos dados (ex: formato datas e telefones).

5. Conclusão

Constituiu objetivo deste projeto a identificação e correção dos problemas de qualidade de dados e enriquecimento dos atributos de morada da informação caracterizadora dos clientes e dos pontos de venda.

Para atingir o objetivo a FinanceQ adquiriu o *software SAS Dataflux* que disponibiliza componentes que permitem identificar e aplicar técnicas de MQD e contratou os serviços da SIGQ para realizar o projeto. No projeto seguiu-se a metodologia *SAS Dataflux* que é composta por três fases: planeamento; ação; e monitorização. Apesar da metodologia aplicada ser semelhante à metodologia mais aclamada na literatura (GTQD (Wang, R.,1998)) é de notar que a metodologia *SAS Dataflux* está mais orientada para projetos. Esta afirmação advém do facto dos primeiros passos após a identificação do âmbito é a análise dos problemas de qualidade nos dados e de ser possível verificar que existe uma fase de tentativa e erro iterativa (monitorização) que culmina quando os problemas identificados são eliminados ou diminuídos.

O processo de limpeza de dados no projeto seguiu os trâmites indicados na metodologia SAS que é semelhante a literatura. Na primeira etapa foi utilizada a componente data *profiling* (Naumann, 2014) para auditar os dados. Utilizando como auxiliar a análise de PQD realizada por Oliveira et al. (2005), de forma sucinta, a análise DP permitiu identificar a existência dos seguintes problemas: valores sem significado; valores a *null*; padrões inadequados para o atributo; erros ortográficos; existência de sinónimos; e valores fora do domínio. Para os atributos de documentação o foco centrou-se em validar o *check digit* para verificar se os documentos eram válidos.

Importa considerar que alguns dos erros identificados não foram identificados de forma explícita na literatura (Oliveira et al., 2005), como por exemplo, o problema de padrões inadequados para o atributo.

Na fase seguinte foram desenvolvidas e implementadas técnicas reativas (Batini et al., 2009) para colmatar os problemas identificados na análise da primeira etapa, sendo as mais aconselhadas quando se pretende corrigir os valores dos dados existentes (Batini et al., 2009). As técnicas estão na sua maioria em consonância com a literatura (Batini et al.,2009; Barateiro & Galhardas, 2005; Batini e Scannapieco, 2006), tendo sido aplicadas técnicas que se enquadram como normalização e enriquecimento de dados. Neste sentido foram designadas regras para validar o formato dos valores; regras para validar a consistência dos atributos; medidas para enriquecer os valores a *null* quando possível; regras para validar os valores das tabelas de referência; regras para validar o *check digit*; medidas para eliminar os valores sem significado; e medidas de normalização para corrigir os erros ortográficos e eliminar a existência de sinónimos.

No final da implementação das regras definidas procedeu-se à atividade de avaliação do processo de melhoria de qualidade. A avaliação foi realizada comparando os valores do relatório DP do início do projeto com o relatório DP efetuado no final projeto e através de testes realizados pelo cliente.

Nesta fase verificou-se que as medidas implementadas permitiram designar corretamente os géneros, sendo que 0,85% dos registos apresentavam erradamente a designação referente ao género masculino, e identificar e corrigir corretamente os títulos de cliente. As ações de limpeza e correção de erros ortográficos foram um sucesso e permitiram eliminar os valores sem significado e corrigir os valores nos atributos, como por exemplo nos atributos de email e *site*. Nos números de telefone as regras implementadas permitiram reorganizar os valores de telefone consoante as prioridades definidas pelo cliente. No atributo profissão a regra implementada permitiu eliminar a existência de sinónimos. Os PQD de valores fora do domínio no atributo nome referiam-se aos títulos de cliente, e após a aplicação das regras criadas verificou-se que os valores ficaram corrigidos e foram utilizados para enriquecer o

atributo título de cliente. Nas moradas o processo de enriquecimento permitiu normalizar os atributos de morada e enriquecer os atributos de código postal.

O processo de melhoria de qualidade de dados permitiu evidenciar a importância da inclusão de validações nos dados como valores e padrões possíveis, tal como a identificação de campos de preenchimento obrigatório ou prioritário. A necessidade destas validações ficou percecionada nos atributos de profissão, título de cliente e datas.

Para além das medidas de validação, este processo salientou a importância do cuidado na inserção dos valores nos respetivos campos e da importância de informar o utilizador do correto preenchimento, para que não haja problemas de valores fora do domínio.

As regras implementadas na sua generalidade tiveram como objetivo melhorar a exatidão e objetividade dos valores (ex: correção dos erros ortográficos); tornar os dados mais completos (ex: enriquecimento das moradas); e melhorar a consistência dos dados (ex: formato datas e telefones).

No que respeita a limitações é de salientar o facto de não ter sido possível eliminar todos os PQD, como por exemplo, os valores a *null*, evidenciados nos atributos email, *site* e/ou números de telefone. Apesar do processo de enriquecimento ter alcançado os valores necessários para ser considerado bem-sucedido, devido à complexidade das regras a criar para que houvesse correspondência, não foi possível garantir 100% de registos preenchidos. Outro fator de limitação é o facto de não ter sido possível eliminar por completo a existência de sinónimos no atributo de profissão, uma vez que o trabalho de eliminação efetuado foi completamente manual.

Numa perspetiva de trabalhos futuros seria interessante estudar o trabalho de qualidade efetuado ao nível dos processos que criam os dados, para garantir que os PQD identificados não voltem a acontecer. Seria também interessante alargar o trabalho de qualidade às tabelas que contém informação operacional, como contratos e movimentos, de forma a verificar a extensão dos problemas de qualidade de dados.

6. Bibliografia

Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., & Zhao, J. L. (2016). Transformational issues of big data and analytics in networked business. *MIS Quarterly*, 807-818

Ballou, D. P., & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*, *31*(2), 150–162.

Ballou, D., Wang, R., Pazer, H., & Tayi, G. K. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, *44*(4), 462–484.

Barateiro, J., & Galhardas, H. (2005). A Survey of Data Quality Tools. *Datenbank-Spektrum*, *14*, 15–21.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, *41*(3), 16:1–16:52.

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. *Springer*.

Brodie, M. L. (1980). Data Quality in Information Systems. *Information & Management*, *3*(6), 245–258.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data To Big Impact. *Mis Quarterly*, *36*(4), 1165–1188.

Davenport, T. H., Harris, J. G., & Morison, R. (2010). *Analytics at Work: Smarter Decisions, Better Results*. *Harvard Business School Press Books*.

Huh, Y. U., Keller, F. R., Redman, T. C., & Watkins, A. R. (1990). Data Quality. *Information and Software Technology*, *32*(8), 559–565.

Isaías, P., & Coelho, F. (2013). Web 2.0 Tools Adoption Model. *International Journal of Information Communication Technologies and Human Development*, *5*(3), 64–79.

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D. (2003). A Taxonomy of Dirty Data. *Data Mining and Knowledge Discovery*, *7*(1), 81–99.

LEE, Y. W. (2003). Crafting Rules: Context-Reflective Data Quality Problem Solving. *Journal of Management Information Systems*, *20*(3), 93–119.

Martinez-Lorente, A. R., Dewhurst, F., & Dale, B. G. (1998). Total quality management: origins and evolution of the term. *The TQM Magazine*, *10*(5), 378–386.

Müller, H., & Freytag, J. (2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing. *Technical Report,* HUB-IB-164,Humboldt University Berlin , 1–23.

Naumann, F. (2014). Data profiling revisited. *ACM SIGMOD Record*, *42*(4), 40–49.

Oliveira, P., Rodrigues, F., Henriques, P., & Galhardas, H. (2005). A Taxonomy of Data Quality Problems. Journal of Data and Information Quality - JDIQ .

Orr, K. (1998). Data quality and systems theory. *Communications of the ACM*, *41*(2), 66–71.

Porter, Michael E; Heppelmann, J. E. (2014). Managing the Internet of things. *Harvard Business Review*, (November), 65–88.

Queen, M. K. (2016). How to Find Your Perfect Match Using SAS ® Data Management, 1–10.

Rahm, E., & Do, H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3–13.

Raman, V., & Hellerstein, J. M. (2001). Potter's Wheel: An Interactive Data Cleaning System. *Vldb*, *01*, 381–390.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), 79–82.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110.

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, *41*(2), 54–57.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95.

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), 58–65.

Wang, R. Y., Kon, H. B., & Madnick, S. E. (1993). Data Quality Requirements Analysis and Modeling. *Data Engineering*, *8*(April), 670–677.

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

7. Anexos

Anexo I - Métricas e Análises aplicadas pela ferramenta de *data profiling*

Métricas/Análises	Descrição
Posição Ordinal	Posição ou número de ordem do atributo na tabela
Contagem	Número de registos em análise
Contagem Valores Null	Número de registos que está em falta
Percentagem Valores Null	Percentagem de registos que está em falta
Contagem Valores em Branco	Número de registos que apresentam o atributo em análise em branco
Valor Mínimo	Valor mínimo do atributo em análise
Valor Máximo	Valor máximo do atributo em análise
Moda	Moda do atributo em análise
Contagem de Padrão	Contagem do número de padrões distintos do atributo em análise
Contagem Valores Únicos	Contagem do número de registos únicos
Unicidade	Percentagem de valores únicos no atributo em análise
Chave Primária Candidata	Candidato a chave primária do atributo
Tipo de Dado	Tipo de dados presentes no atributo em análise
Tamanho dos dados	Tamanho do atributo em análise
Tipo de Dado Real	Metadata identificada
Tamanho Mínimo	Tamanho do menor valor no atributo em análise
Tamanho Máximo	Tamanho do maior valor no atributo em análise
Média	Média da população/amostra no atributo em análise
Mediana	Mediana da população/amostra no atributo em análise
Contagem Valores Dif. Null	Contagem de registos não nulos
Nullable	Admite valores em falta
Número de Casas Decimais	Número de casas decimais do atributo em análise
Desvio Padrão	Desvio Padrão da população/amostra no atributo em análise
Erro Padrão	Erro padrão dos valores do atributo em análise
Análise de Frequência	Análise e apresentação de valores únicos no atributo em análise
Análise de Padrões	Análise e apresentação dos padrões presentes no atributo em análise
Percentis	Divisão dos dados em 100 partes cada uma com uma percentagem aproximadamente igual de dados
Outliers	Valores que representam os extremos (máximos e mínimos) da população/amostra

Anexo II - Taxonomia de Problemas de Qualidade de Dados

Nível PQD	PQD	Definição
O valor de um atributo de um único tuplo	Valor em falta	Falta de valores em atributos em que se requer preenchimento
	Violação da sintaxe	Discrepância entre a sintaxe definida para o atributo e a que está no valor do atributo
	Valor desatualizado	O valor no atributo não corresponde a situação atual
	Violação do intervalo	Violação do intervalo de valores válidos de um atributo do tipo numérico
	Violação do domínio	Violação dos valores que o atributo pode conter
	Erro ortográfico	Atributos com erros ortográficos
	Valor inadequado para o contexto	O valor do atributo não se adequa ao atributo em questão, mas sim, noutro atributo
	Valor fora do domínio	Múltiplos valores armazenados no mesmo atributo. Alguns atributos deveriam estar noutro atributo
	Valor sem significado	O valor não faz sentido para o contexto do atributo.
	Valor impreciso ou com vários significados	Consequência de utilizar abreviações ou acrónimos nos atributos textuais. Pode levar a diferentes interpretações
	Violação da restrição do domínio	Violação de uma restrição relacionada com o atributo.
Os valores de um único atributo	Violação do valor único	Dois ou mais tuplos referentes a diferentes entidades contêm o mesmo valor num atributo que deveria ser único
	Existência de sinónimos	Uso arbitrário de valores diferentes que são sinónimos no mesmo atributo.
	Violação da restrição do domínio	Violação da restrição associada aos valores que atributo pode assumir em vários tuplos.
Os valores dos atributos de um único tuplo	Tuplo parcialmente vazio	Acontece quando grande parte dos atributos do tuplo não estão preenchidos
	Inconsistência entre valores dos atributos	Violação da dependência entre os valores dos atributos do tuplo
	Violação de restrição	Violação de restrições que envolvem dois ou mais atributos do tuplo
Os valores dos atributos de vários tuplos	Redundância entre entidades	A mesma entidade é representada pela mesma ou equivalente representação em mais do que um tuplo
	Inconsistência entre entidades	Inconsistências ou contradições entre os valores de um ou mais atributos referentes a mesma entidade em diferentes tuplos
	Violação da restrição do domínio	Violação da restrição relacionada com a tabela
Relacionamentos entre múltiplas relações	Violação dependência funcional	Tuplo que contém valor referente a uma chave estrangeira que não existe como chave primária na respetiva tabela

	Referência desatualizada	Apesar da integridade referencial estar a ser respeitada, o valor referente a chave estrangeira não está atualizada
	Inconsistência de sintaxe	O mesmo atributo representado em diferentes tabelas apresenta sintaxes diferentes
	Inconsistência entre atributos relacionados	Inconsistências entre valores de atributos em tabelas em que existe relacionamento entre os atributos
	Referências circulares entre tuplos	Existência de referências em atributos de diferentes tabelas, mas relacionados entre si, que criam um ciclo sem sentido
	Violação da restrição do domínio	Violação da restrição que envolve relacionamentos de múltiplas tabelas
Múltiplas fontes de dados	Inconsistência de sintaxe	Atributos do mesmo tipo, mas de diferentes fontes, apresentam sintaxes diferentes.
	Diferentes unidades de medida	Atributos relacionados, mas de diferentes fontes, apresentam unidades de medida diferentes
	Representação inconsistente	O mesmo tipo de atributo de diferentes fontes é representado com conjunto de de valores diferentes.
	Diferentes níveis de agregação	O nível de detalhe apresentado em tabelas equivalentes de diferentes fontes não é igual
	Existência de sinónimos	Uso arbitrário de valores diferentes que são sinónimos em atributos equivalentes de diferentes fontes
	Existência de homónimos	Utilização de valores homónimos em atributos equivalentes de diferentes fontes
	Redundância entre entidades	A mesma entidade é representada pela mesma ou equivalente representação em mais do que um tuplo de diferentes fontes
	Inconsistência entre entidades	Inconsistências ou contradições entre os valores de um ou mais atributos referentes a mesma entidade em diferentes tuplos de diferentes fontes
	Violação de restrição	Relações equivalentes em diferentes fontes, individualmente respeitam a restrição, mas violam quando integradas como um todo.

Fonte : Adaptado de Oliveira P. et al. (2005)

Anexo III - Atributos Caracterizadores da Tabela Clientes

TIPO DE ATRIBUTO	ATRIBUTO	DESCRIÇÃO
DATA	DATABR	Data de Abertura
DATA	DATANI	Data de Nascimento
DATA	EMPDAT	Data de Contratação
DATA	DTAVBI	Data de Validade do BI
DATA	HABDAT	Habita Desde
DOCUMENTO	NUMDOC	Número de documentação
DOCUMENTO	NUMCONT	NIF
DOCUMENTO	NPCNUM	NIF Coletivo
DOCUMENTO	NUMBI	BI
DOCUMENTO	NUMCARTA	Número Carta de Condução
DOCUMENTO	NUMNIB	NIB
GÊNERO	CLISEX	Género
TITULO CLIENTE	CLITIT	Título do Cliente
NOME CLIENTE	CLINOM	Nome
TIPO CLIENTE	TIPCLI	Tipo de Cliente
DESIGNAÇÃO TRABALHO	TITTRA	Título de Trabalho
TELEFONE	TELCASA	Telefone de Casa
TELEFONE	TELFAX	Fax
TELEFONE	TELMOV	Telemóvel
TELEFONE	TELMTRB	Telemóvel de Trabalho
TELEFONE	TELTRBO	Telefone de Trabalho
EMAIL	EMAIL	Email
EMAIL	EMAILT	Email de Trabalho
MORADA	NUMRUA	Número da Rua
MORADA	MORADA	Morada
MORADA	PORTA	Número da Porta
MORADA	ANDAR	Andar
MORADA	LOCALID	Localidade
MORADA	CODPOS	Código-Postal

Anexo IV - Atributos Caracterizadores da Tabela de Pontos de Venda

TIPO DE ATRIBUTO	ATRIBUTO	DESCRIÇÃO
NOME	VNDNOME	Nome
DESIGNAÇÃO COMERCIAL	VNDDCOM	Designação Comercial
GÊNERO	VNDGEN	Género
DOCUMENTO	VNDNFC	NIF
EMAIL	VNDMAIL	Email
SITE	VNDINTER	Site
DATA	VNDCDCTCO	Cod. Certidão Comercial
DATA	VNDDABR	Data de Abertura
DATA	VNDDALT	Data de Alteração/Confirmação
DATA	VNDDANU	Data de Anulação
DATA	VNDLCRGD	Data de alteração da linha Crédito Global
DATA	VNDLCRUD	Data de alteração da linha Crédito Unsecure
DOCUMENTO	VNDNIB1	NIB
DOCUMENTO	VNDNIB2	NIB
DOCUMENTO	VNDNIB3	NIB
DOCUMENTO	VNDNIB4	NIB
DOCUMENTO	VNDNIBC	NIB
DOCUMENTO	VNDNIBF	NIB
DOCUMENTO	VNDNIBFS	NIB
DOCUMENTO	VNDNIBOS	NIB
DOCUMENTO	VNDNIBPS	NIB
DOCUMENTO	VNDNIBR	NIB

Anexo V - Exemplo de Utilização de Linguagem EEL